An Evaluation of the UIT-VSFC Dataset Using Modern Machine Learning Techniques and Word Embeddings

Quỳnh Dương Vũ Xuân
School of Industrial Engineering and Management
International University - Vietnam National University
Ho Chi Minh City
Ho Chi Minh City, Vietnam
dvxquynh@hcmiu.edu.vn

Kanjana Laosen
AI-TaSI Center
College of Computing
Prince of Songkla University
Phuket Campus
Phuket, Thailand
kanjana.l@phuket.psu.ac.th

Nasith Laosen
Department of Digital Technology
Faculty of Science and Technology
Phuket Rajabhat University
Phuket, Thailand
nasith.l@pkru.ac.th

2021 25th International Computer Science and Engineering Conference (ICSEC) | 978-1-6654-1197-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSEC53205.2021.9684597

Abstract—**Feedback from students during the course and upon course completion has become a powerful resource to improve teaching quality and enhance student's learning experience. However, the available data for free use is limited, especially for a low-resource language like Vietnamese. Currently, there is only one dataset in the education domain, called the Vietnamese Students' Feedback Corpus (UIT-VSFC), that has been published for free use. This study therefore aims at evaluating the available corpus to use as a benchmarking dataset for conducting future researches as well as developing real-world applications. In this paper, deep neural network (DNN) and recurrent neural network (RNN) models are developed employing a word embedding method for two different tasks, i.e., topic and polarity classification. The experimental results show that DNN models outperform RNN models with 85.22% (>84.30%) and 88.56% (>86.32%) of accuracy for topic and polarity classification, respectively. Error analysis is conducted to explore the confusion of labeling and imbalance of data in the dataset. Workarounds for solving the problems are presented together with their results.**

Index Terms—**aspect-based sentiment analysis, education domain, student's feedback, Vietnamese**

I. Introduction

There is no doubt that education plays a crucial role in the development of any nation. To obtain a better education quality, academic institutions need to know which aspects (e.g., curriculums, lecturers, or facilities) of them that need to be improved. Students' feedback is considered as a powerful data source to identify such aspects as it reflects a comprehensive picture of the institutions. Students' feedback is usually given in the form of textual description. This allows aspect-based sentiment analysis to take place.

Aspect-based sentiment analysis is a process of identifying aspects and their emotional polarity (e.g., positive, negative, and neutral) from texts. It attracts many researchers around the world and has been widely applied to various domains including financial market domain [1], customer relationship domain [2], product service domain [3] and education domain [4]–[9]. However, the idea is still fresh in developing countries like Vietnam as the fact that Vietnamese is a resource-poor language, especially in the education domain.

To the best of our knowledge, there is only one Vietnamese language dataset in the education domain that was published for free access, i.e., the Vietnamese Students' Feedback Corpus (UIT-VSFC). Such dataset was proposed in [10] and was evaluated using two traditional classification techniques, i.e., Naive Bayes and Maximum Entropy. The experiments in [10] showed that Maximum Entropy achieved promising classification results.

Although the traditional classification techniques are still applicable, modern classification techniques, e.g., deep neural networks and deep learning, are currently found to be popular and efficient for sentiment analysis. Moreover, since the Vietnamese data resource in the education domain is limited, the UIT-VSFC dataset has a high chance to be used as a benchmark dataset for research or used for developing real-world applications. Therefore, for all the above reasons, we believe that evaluating the UIT-VSFC dataset with modern techniques is necessary. In this paper, deep neural networks and recurrent neural networks together with a word embedding technique (i.e., Word2Vec) are used to evaluate the dataset. Based on the evaluation results, error analysis is conducted to closely examine and identify shortcomings of the dataset. Lastly, workarounds for solving the problems are provided with results.

Our contributions could be summarized as four main points. First, we provide a different approach for solving aspect-based sentiment analysis benchmarking the available data resource. Secondly, we conduct error analysis to better understand the pros and cons of the dataset for further improvement. Thirdly, we enhance the performance of the classification. Lastly, the findings of this paper can be used as a guideline for researchers and developers who are conducting research or developing real-world applications using this dataset.

978-1-6654-1197-4/21/$31.00 ©2021 IEEE

The rest of this paper is organized as follows: Section II reviews related works. Section III describes the UIT-VSFC dataset. Section IV presents experiments and results. Section V analyzes errors and gives workaround results. Section VI provides conclusions and future work.

II. Related Works

In recent years, the education domain has attracted the attention of the research community via sentiment analysis and opinion mining. There are various studies that have been conducted for the English language. In particular, Sindhu et al. [5] proposed a supervised aspect-based opinion mining system based on deep learning models. Z. Kastrati et al. [6] proposed a framework to analyze opinions of students that are expressed in reviews. G. S. Chauhan et al. [7] gathered posted feedback from students and teachers, and employed a lexicon-based approach to evaluate the teaching–learning process. M. Moreno-Marcos et al. [8] applied different machine learning algorithms via their studying to check how the results can provide information about learners' emotions or patterns. R. Bogdan et al. [9] conducted an English survey exploring how Web 2.0 technologies can be applied into teaching embedded systems courses.

In the last few years, the Vietnamese education domain has started to attract the attention of the local research community as the demand to discover learners' insight as well as enhance teaching programs is increasing. However, the number of research works in this domain is still small. In 2018, K. V. Nguyen et al. [10] proposed a free corpus in the education domain (the UIT-VSFC dataset). The Maximum Entropy classifier was used in [10] to evaluate the proposed dataset and it achieved 84.03% and 87.94% of the overall F1-score of topic and polarity classification tasks, respectively. In 2020, T. P. G. Nguyen et al. [11] proposed a simple system to categorize students' feedback into positive, negative, and neutral with the highest accuracy reaching 91.36% using the Maximum Entropy classifier. Nevertheless, the dataset used in [11] was not published for public. From the literature, our research is one of the few attempts that use deep learning/deep neural networks together with a word embedding technique for sentiment analysis in the Vietnamese education domain.

III. Data Resource Description

The UIT-VSFC dataset consists of more than 16,000 sentences, each of which is annotated for two tasks, i.e., topic and polarity classification tasks.

Regarding the topic classification task, four topics are considered, i.e., LECTURER, CURRICULUM, FACILITY, and OTHER. In particular, sentences which show feedback towards teaching activities, lecturers, knowledge, or attitude-toward-students are annotated as LECTURER, whereas sentences that give justification for a program's quality, knowledge, assignments, or grading are annotated as CURRICULUM. FACILITY, on the other hand, are sentences that give feedback regarding the quality of facilities in the university (e.g., computers, Wi-Fi, air conditioners, and light). Lastly, sentences belonging to OTHER do not contain the above mentioned topics.

Regarding the polarity classification task, three sentiment polarities are considered, i.e., POSITIVE, NEGATIVE, and NEUTRAL. Particularly, POSITIVE sentences express their positive emotion or compliments towards one of the topics under consideration (LECTURER, CURRICULUM, FACILITY, and OTHER), whereas NEGATIVE sentences illustrate the drawbacks or disadvantages regarding one of the topics. Sentences that do not contain the above mentioned sentiment would be annotated as NEUTRAL.

For example, the sentences (i) "cô vui tính, nhiệt tình" (the female lecturer is humorous and enthusiastic), (ii) "giáo trình chưa có hợp lý" (the curriculum is not appropriate), (iii) "phòng máy, thiết bị cũ" (machine room and equipment are old), and (iv) "cám ơn cô đã dạy lớp em" (thank you for teaching my class), are annotated as LECTURER#POSITIVE, CURRICULUM#NEGATIVE, FACILITY#NEGATIVE and OTHER#NEUTRAL, respectively.

The UIT-VSFC dataset is divided into three sets, i.e., training set, validation set, and testing sets. Training set occupied around 70% of the total dataset, whereas validation and test sets are around 20% and 10%, respectively.

IV. Experiments and Results

A. Experiment Settings

1) Data Preparation: As mentioned in Section I, we try to evaluate the UIT-VFSC dataset by using deep neural network (DNN) and recurrent neural network (RNN) models. Since those models cannot process textual data directly, we need to convert the textual data into some numerical form. In order to do that, the Word2Vec technique [12] is employed to create vectors for representing words appearing in the dataset. These vectors are also called word embeddings. In this paper, each word in the dataset is represented by a 1000-element vector, as shown in Fig. 1. The obtained vectors will be used as input of the classification models.

The key power of the Word2Vec technique is that words with similar meaning and context will appear close to each other on a vector space, whereas words with different meaning and context appear far away from each other. Fig. 2 shows resulting vectors (converted from 1000-dimensional to 2-dimensional) created by the Word2Vec technique. As shown in the figure, words are grouped into several categories including "teaching verbs", "attitude-towards-students related words", "teaching feedback", "teaching habits", "lecturer's characteristics related-words", and "nouns related words". The words in each category have similar meaning and context. For instance, (i) the words "thầy" (male lecturer) and "cô" (female lecturer) are grouped into the "nouns related words" category and (ii) the words "vui tính" (humorous), "nhiệt tình" (enthusiastic), and "giúp đỡ" (helpful) are grouped into the "lecturer's characteristics related words" category.

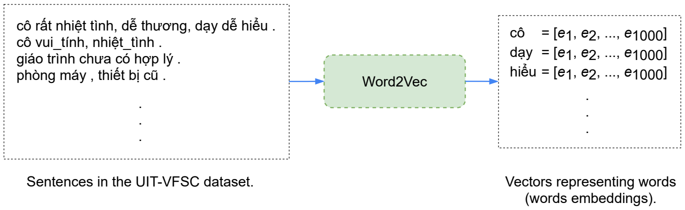

cô rất nhiệt tình, dễ thương, dạy dễ hiểu .
cô vui_tính, nhiệt_tình .
giáo trình chưa có hợp lý .
phòng máy , thiết bị cũ .
Word2Vec
cô = [e_1, e_2, ..., e_{1000}]
dạy = [e_1, e_2, ..., e_{1000}]
hiểu = [e_1, e_2, ..., e_{1000}]
Sentences in the UIT-VFSC dataset.
Vectors representing words (words embeddings).

Fig. 1. Creating vectors representing words in the dataset.

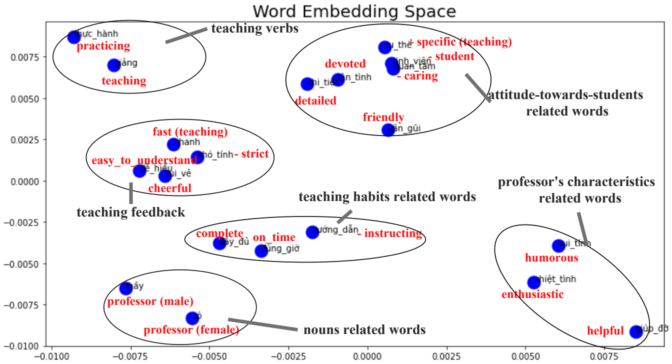
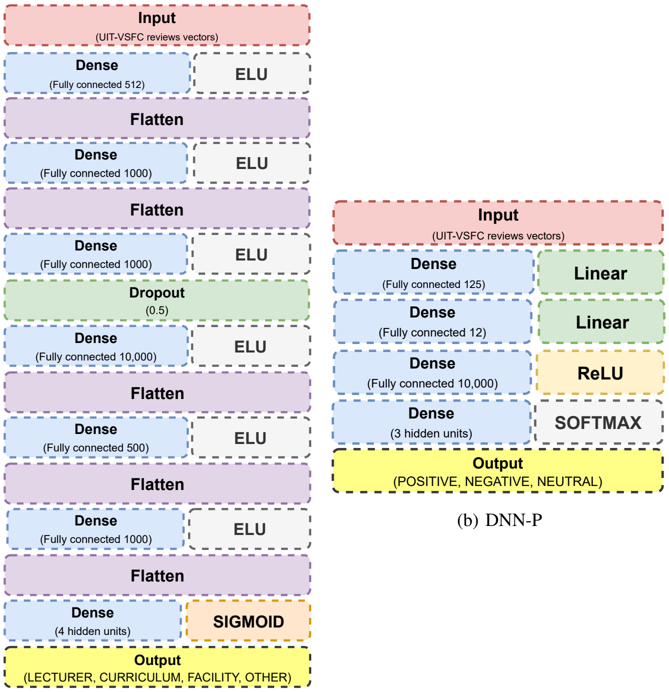


Fig. 2. Plotting word vectors on a 2-dimensional space.



(b) DNN-P

(a) DNN-T

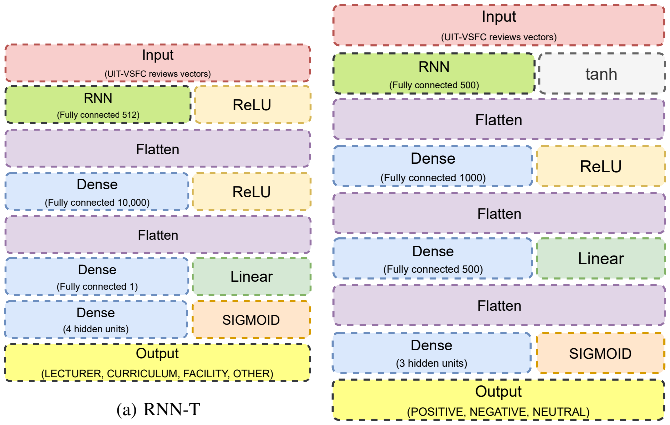
Fig. 3. Deep neural network models for topic and polarity classification.

2) *Architectures of Classification Models:* As mentioned in Section III, the UIT-VSFC dataset is annotated for two tasks, i.e., topic and polarity classification. The DNN models in Fig. 3(a) and Fig. 3(b), referred to as DNN-T and DNN-P, are used for topic and polarity classification, respectively. In the same manner, the RNN models in Fig. 4(a) and Fig. 4(b), referred to as RNN-T and RNN-P, are used for topic and polarity classification, respectively. The Tensorflow library and the Adam learning algorithm are used to train all models.

B. Experimental Results

We trained and validated the DNN-T, RNN-T, DNN-P, and RNN-P by using the training set and the validation set, respectively (cf. Section III). Then the models were evaluated by using the testing set. The classification results were compared to those of the Maximum Entropy classifier (referred to as Max-Ent) originally reported in [10]. The comparisons are described below.

1) *Topic Classification:* Table I shows the precision, recall, and F1-score of Max-Ent, RNN-T, and DNN-T. Although the performance of DNN-T is not higher than Max-Ent in all cells, the F1-score of DNN-T achieves better value in all topics. Particularly, the classes LECTURER, CURRICULUM, FACILITY, and OTHER obtain values of 91.59%, 68.89%, 89.51%, and 41.53%, which are higher than those of Max-Ent with 91.12%, 67.19%, 88.73%, and 38.10%, respectively. The overall accuracy for DNN-T is more than 85%. On the other hand, RNN-T wins only the precision score of OTHER and Recall score of LECTURER topics with 70.37% and 92.92%, respectively. Interestingly, DNN-T appears to be quite promising model with higher performance compared to Max-Ent and RNN-T. Note that the accuracy of Max-Ent was not reported in [10].



(a) RNN-T

(b) RNN-P

Fig. 4. Recurrent neural network models for topic and polarity classification.

2) *Polarity Classification:* Table II shows the performance of polarity classification of the Max-Ent, RNN-P, and DNN-P. Although the difference in performance of DNN-P and Max-ent is small, more cells have better scores using Max-Ent. Interestingly, the class NEUTRAL gets 58.33% of precision using DNN-P which is higher than that of Max-Ent by more than 8%. However, the F1-score of the NEUTRAL class is still 3% less than that of Max-Ent. In contrast, RNN-P gets lower performance in all cells with an overall accuracy of 86.32%. In general, DNN-P achieves better performance compared to RNN-P with more than 88% of accuracy.

TABLE I
COMPARING TOPIC CLASSIFICATION RESULTS (%)

	Precision			Recall			F1-score		
	Max-Ent	RNN-T	DNN-T	Max-Ent	RNN-T	DNN-T	Max-Ent	RNN-T	DNN-T
LECTURER	90.17	88.81	**91.61**	92.10	**92.92**	91.57	91.12	90.08	**91.59**
CURRICULUM	**67.07**	65.81	64.34	67.31	67.30	**74.13**	67.19	66.55	**68.89**
FACILITY	90.65	90.07	**90.78**	86.90	81.37	**88.28**	88.73	85.50	**89.51**
OTHER	45.61	**70.37**	63.64	**32.70**	23.89	30.82	38.10	35.68	**41.53**
Accuracy								84.30	**85.22**

TABLE II
COMPARING POLARITY CLASSIFICATION RESULTS (%)

	Precision			Recall			F1-score		
	Max-Ent	RNN-P	DNN-P	Max-Ent	RNN-P	DNN-P	Max-Ent	RNN-P	DNN-P
POSITIVE	**91.69**	90.17	92.25	**90.94**	88.30	90.69	91.32	89.23	**91.41**
NEGATIVE	**87.69**	84.42	86.00	93.54	91.55	**94.18**	**90.52**	87.84	89.9
NEUTRAL	50.00	48.15	**58.33**	**25.75**	23.35	20.95	**33.99**	31.45	30.83
Accuracy								86.32	**88.56**

V. ERROR ANALYSIS

After we evaluated the dataset, we found that the classes CURRICULUM and OTHER in the topic classification, and the class NEUTRAL in the polarity classification got lower precision, recall, and F1-score compared to the other classes (cf. Tables I and II). Therefore, to increase the performance of the classification, we analyzed the results, identified potential sources of errors, and worked around to solve the problems. The potential sources of errors and results of workarounds are described below.

A. Potential Sources of Errors

1) Confusion of Labeling: Regarding the topic classification, as shown in Table III, the classification errors mainly occur between the classes CURRICULUM and LECTURER, i.e., 164 sentences of CURRICULUM are wrongly classified as LECTURER. We then examined the training set closely and found that there are some sentences that express the curriculum's content but they are labelled as LECTURER. For example, "nội dung môn học có phần thiếu trọng tâm, hầu như là chung chung, khái quát khiến sinh viên rất khó nắm được nội dung môn học" (the subject content is somehow lacking in focus, almost in general, making it difficult for students to grasp the subject content) or "lượng kiến thức quá nhiều trong một học kỳ" (the amount of knowledge is too much to handle for one semester). These two classes are related by nature and easily get wrongly labeled. This might cause confusion for training and testing the models, which then reduces the overall performance of the models.

2) Imbalance of Data: During the error analysis, we also found that the number of sentences of the classes OTHER and NEUTRAL are small compared to the other classes. In particular, in the training set, the classes OTHER and NEUTRAL have around 500 and 400 sentences, respectively, whereas the total number of training data is more than 11,000 sentences. This causes the imbalance in the dataset and might lead to the low performance of these two classes (cf. Tables I and II).

B. Workarounds and Results

1) A Workaround for Labeling Confusion: In order to eliminate confusion between the classes LECTURER and CURRICULUM, we reassigned CURRICULUM sentences to LECTURER if they consists of LECTURER related words, such as: "thầy" (male lecturer), "cô" (female lecturer), "nhiệt tình" (enthusiastic), and so on. For example, the sentence "bài tập cô cho nhiều nên làm hơi đuối" (the homework the female lecturer assigned is too much, feeling a bit exhausted), which was originally labeled as CURRICULUM, was reassigned to LECTURER. On the other hand, we also reassigned sentences consist of CURRICULUM related words originally labeled as LECTURER to CURRICULUM. For example, the sentences "chương trình dạy lý thuyết và thực hành nên khớp với nhau hơn, tránh tình trạng thực hành trước khi học lý thuyết" (the theory and practice curriculum should be more closely matched, avoiding the situation of practicing before learning the theory) and "có một số phần chưa phân bổ hợp lý lắm" (some parts are not properly distributed) were reassigned to CURRICULUM.

After we reconsidered training and testing datasets, we achieved an improvement in both RNN-T and DNN-T compared to using the original datasets. Table IV shows the performance of RNN-T and DNN-T after the dataset has been improved. As we can see from the table, DNN-T achieves higher scores compared to RNN in most of the cells. DNN-T also outperforms RNN-T for the overall performance with 87.08% compared to 85.31%. On the other hand, if we compare this result with the result achieved using the original dataset in Section IV, RNN-T has improved its performance from 84.30% to 85.31%. Similarly, DNN-T increases from 85.22% to 87.08%. Table V shows the confusion matrix after improving the dataset. We might conclude that the improved dataset has helped to enhance the performance of both models.

2) A Workaround for Imbalanced Data: In an attempt to reduce the impact of imbalanced data and increase the performance of the classification of the classes OTHER and NEUTRAL, the following steps were performed:

TABLE III
THE CONFUSION MATRIX OF THE DNN-T MODEL ON THE TESTING SET

	Topic	Predicted class				Total
		Lecturer	Curriculum	Facility	Other	
Actual class	Lecturer	2,128	**136**	5	21	2,285
	Curriculum	**164**	397	5	6	572
	Facility	6	18	121	0	145
	Other	75	38	3	43	159

TABLE IV
COMPARING TOPIC CLASSIFICATION RESULTS AFTER IMPROVING THE DATASET (%)

	Precision		Recall		F1-score	
	RNN-T	DNN-T	RNN-T	DNN-T	RNN-T	DNN-T
Lecturer	90.62	**90.98**	92.62	**94.48**	91.61	**92.69**
Curriculum	71.10	**73.88**	72.17	**74.00**	71.63	**73.94**
Facility	**90.77**	89.44	82.52	**88.81**	86.45	**89.12**
Other	47.37	**65.75**	**33.75**	30.00	39.42	**41.20**
Accuracy					85.31	**87.08**

TABLE V
THE CONFUSION MATRIX OF THE DNN-T MODEL ON THE TESTING SET AFTER IMPROVING THE DATASET

	Topic	Predicted class				Total
		Lecturer	Curriculum	Facility	Other	
Actual class	Lecturer	2,138	**102**	6	17	2,263
	Curriculum	**145**	444	5	6	600
	Facility	5	9	127	2	143
	Other	62	46	3	48	159

TABLE VI
TOPIC CLASSIFICATION RESULTS AFTER RESAMPLING TRAINING SET (%)

	Precision		Recall		F1-score	
	RNN-T	DNN-T	RNN-T	DNN-T	RNN-T	DNN-T
Lecturer	78.00	73.92	72.61	73.14	75.21	73.53
Curriculum	77.36	76.15	78.85	78.15	78.10	77.14
Facility	86.01	85.71	84.83	82.76	85.42	84.21
Other	50.00	49.02	55.35	47.17	**52.54**	48.08
Accuracy					74.68	73.24

TABLE VII
POLARITY CLASSIFICATION RESULTS AFTER RESAMPLING TRAINING SET (%)

	Precision		Recall		F1-score	
	RNN-T	DNN-T	RNN-T	DNN-T	RNN-T	DNN-T
Positive	73.59	71.99	88.81	88.80	80.49	79.52
Negative	91.24	90.12	78.32	78.94	84.29	84.16
Neutral	44.44	45.20	53.15	46.15	**48.41**	45.67
Accuracy					78.85	78.47

- *Step 1:* Resample the topic and polarity training sets to reduce the number of sentences of each majority class to 2,000 sentences, while keeping the number of sentences in the minority classes.
- *Step 2:* Shuffle the resample training sets several times to change their orders.
- *Step 3:* Use Word2Vec to create word embeddings.
- *Step 4:* Apply the synthetic minority oversampling technique (SMOTE) [13] to generate instances of the minority classes to be equal to the majority classes with 2,000 samples.
- *Step 5:* Use the new training sets obtained from Step 4 and the original testing sets to train and test the models.

The topic and polarity classification results are shown in Tables VI and VII, respectively. Compared to Tables I and II, the classes OTHER and NEUTRAL improve their F1-scores from 41.53% to 52.54% and from 33.99% to 48.41%, respectively. However, other majority classes in both topic and polarity classification reduce their performance, which might be a result of reducing their training sample size. Based on the result of this experiment, the SMOTE technique might not be the best choice for dealing with the imbalance of data of this dataset. Other techniques should be further investigated for further improvement.

VI. Conclusion

In this study, deep neural network and recurrent neural network models have been built for aspect-based sentiment analysis in Vietnamese education domain. By applying the Word2Vec technique on the UIT-VSFC dataset, we achieve more than 85% and 88% of accuracy for both topic and polarity classification, respectively. Error analysis has been conducted to identify two shortcomings of the dataset, i.e., confusion of labeling and imbalance of data. Workarounds for solving such problems have been presented and their results have been reported. We also found that imbalance of data still be a challenging problem for researchers and developers who want to use the UIT-VSFC dataset. In the future, we plan to investigate other potential methods for dealing with the imbalance of data to further improve the performance of minority classes as well as developing a university review system based on the UIT-VSFC dataset.

References

[1] A. E. O. Carosia, G. P. Coelho, and A. E. A. Silva, “Analyzing the brazilian financial market through portuguese sentiment analysis in social media,” *Applied Artificial Intelligence*, vol. 34, no. 1, pp. 1–19, 2020.

[2] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, “Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning,” *IEEE Access*, vol. 8, pp. 23 522–23 530, 2020.

[3] N. Capuano, L. Greco, P. Ritrovato, and M. Vento, “Sentiment analysis for customer relationship management: An incremental learning approach,” *Applied Intelligence*, vol. 51, no. 6, pp. 3339–3352, 2021. [Online]. Available: https://doi.org/10.1007/s10489-020-01984-x

[4] V. D. Nguyen, K. V. Nguyen, and N. L.-T. Nguyen, “Variants of long short-term memory for sentiment analysis on vietnamese students’ feedback corpus,” in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 2018, pp. 306–311.

[5] I. Sindhu, S. Muhammad Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, “Aspect-based opinion mining on student’s feedback for faculty teaching performance evaluation,” *IEEE Access*, vol. 7, pp. 108 729–108 741, 2019.

[6] Z. Kastrati, A. S. Imran, and A. Kurti, “Weakly supervised framework for aspect-based sentiment analysis on students’ reviews of moocs,” *IEEE Access*, vol. 8, pp. 106 799–106 810, 2020.

[7] G. S. Chauhan, P. Agrawal, and Y. K. Meena, “Aspect-based sentiment analysis of students’ feedback to improve teaching–learning process,” in *Information and Communication Technology for Intelligent Systems*, S. C. Satapathy and A. Joshi, Eds. Singapore: Springer Singapore, 2019, pp. 259–266.

[8] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, I. Estévez-Ayres, and C. D. Kloos, “Sentiment analysis in moocs: A case study,” in *2018 IEEE Global Engineering Education Conference (EDUCON)*, 2018, pp. 1489–1496.

[9] R. Bogdan, N. Pop, and C. Holotescu, “Using web 2.0 technologies for teaching technical courses,” *AIP Conference Proceedings*, vol. 2071, p. 050003, 01 2019.

[10] K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L.-T. Nguyen, “UIT-VSFC: Vietnamese students’ feedback corpus for sentiment analysis,” in *10th International Conference on Knowledge and Systems Engineering (KSE)*, 2018, pp. 19–24.

[11] N. T. P. Giang, T. T. Dien, and T. T. M. Khoa, “Sentiment analysis for university students’ feedback,” in *Advances in Information and Communication*, K. Arai, S. Kapoor, and R. Bhatia, Eds. Cham: Springer International Publishing, 2020, pp. 55–66.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, Jun 2002. [Online]. Available: http://dx.doi.org/10.1613/jair.953